Cool URIs for the DDC: Towards Web-Scale Accessibility of a Large Classification System

Michael Panzer
OCLC, USA
panzerm@oclc.org

Abstract

The report discusses metadata strategies employed and problems encountered during the first step of transforming the DDC into a Web information resource. It focuses on the process of URI design, with regard to W3C recommendations and Semantic Web paradigms. Special emphasis is placed on usefulness of the URIs for RESTful web services.

Keywords: Dewey Decimal Classification; metadata; Uniform Resource Identifiers; web service architecture; classification systems; World Wide Web; REST

1. Introduction

The Dewey Decimal Classification (DDC)[63] system, if it wants to stay relevant to its present and to embrace future users, will have to face the challenge to build a presence on the (Semantic) Web that is not only actionable, but also convenient and useful to its participants. Existing on the Web is the first and currently most important step to potentially become part of "higher-level Web artifacts" that are being built "out of existing Web parts" (T. V. Raman).

Some advances in putting *bibliographic* data and standards on the Web are indeed visible. WorldCat identifiers (OCLC numbers minted as URIs in the worldcat.org namespace) are forming the basis of globally scoped manifestation identifiers for library material; the Library of Congress has recently added permalinks to its catalog records. With regard to subject authority metadata, however, most initiatives keep a very low profile, despite the fact that terminologies, controlled vocabularies, taxonomies, etc., are among the most valuable (and costly) assets of the library community. The relevance of controlled vocabularies for bibliographic standards has become the focus of recent discussion (Coyle & Hillmann, 2007).

The tools and formats that allow those knowledge organization systems to become part of the Semantic Web are emerging. It is now up to providers to rethink historically grown knowledge organization systems (KOS) in terms of these new technologies and make them available for recombination and reuse.

2. Paradigms of Identification, Location, Access

For a resource to be visible on the Web, the single most important piece of information is its URI. It weaves Web resources into the Semantic Web; it connects "things" with information resources describing them, binds information resources together, and (via http) provides information about their relationships. In short, it provides "scaffolding" as well as acts as a "micro-billboard" (Stuart Weibel) for resources.

A URI (Berners-Lee, Fielding, & Masinter, 2005) is commonly defined as a string of characters used to identify or name a single resource. This definition seems odd given the fact that the architecture of the World Wide Web is mainly concerned with representations of *information* resources. Yet, as it is very useful to assign URIs to things that may not be information resources, the discussion about whether the re-entry of the distinction "information

[63] DDC, Dewey, Dewey Decimal Classification, WebDewey, and WorldCat are registered trademarks of OCLC Online Computer Library Center, Inc.

resource/non-information resource" should be allowed into the system of the Web leads to what is now known as its "identity crisis". It was essentially resolved by simply not allowing this re-entry, making the system/environment distinction, but at the same time leaving the environment as an unmarked state. Therefore, the objects identified by URIs are either information resources or things that may or may not be information resources, i.e.,, more plainly, anything.

Among different URI schemes, choosing http is considered best practice for the Semantic Web, because it "can be resolved by any client without requiring the use of additional plug-ins or client setup configuration" (Berrueta & Phipps, 2008, sec. Naming). (Therefore, the resulting URIs are, in fact, URLs.) An information resource returns representations of the identified resource in response to http requests, a process called dereferencing.

Minting URIs for the DDC is not without complications. While other KOS, mostly thesauri, have been using some kind of internal identification for some time that they now might surface, the situation for the DDC is quite the opposite. It was built from the start upon a set of visible identifiers, the Dewey numbers, which should feature prominently in every URI scheme, even if they need to be augmented considerably to satisfy modern standards of Web architecture.

A naming scheme has to be adopted that both exposes the structure of the DDC for addressability and reference and makes sense to agents (clients) using the Web service by asking questions about DDC resources. To put it a different way: The scheme has to be specific to the DDC as well as adhere to the expectations (i.e.,, standards) of the general and the Semantic Web.

The initial questions are: What taxonomy-level and concept-level metadata elements provided by the DDC should be included in the URI (Mendelsohn & Williams, 2007)? How easy should it be to construct an identifier based on previous classification data, e.g., tag 082 in MARC Bibliographic records? How semantically loaded should they be?

The Web community has quite different approaches when it comes to URI design. Tim Berners-Lee, for example, in his "Axiom of URI opacity", states that URIs must not contain any elements that can be connected to the resource in a meaningful way, as such elements might raise expectations about the representation that may or may not be fulfilled upon dereferencing the URI. Since URIs are often implemented as late-binding, (practically) nothing about the information resource referenced by the URI should be inferred until the identifier is dereferenced and its representation is retrieved (W3C Technical Architecture Group, 2004, sec. 2.5).

This axiom or – rather – best-practice recommendation is meant to discourage the derivation of metadata from general data of unknown status ("sniffing"). Metadata that can be acquired this way is often closely related to the document or representation of the resource rather than the resource itself. In addition, data elements in URIs are categorized as "external reference metadata", which is deemed to be the least authoritative metadata source in the context of Web architecture (Fielding & Jacobs, 2006). This type of metadata might depend on not only the intrinsic characteristics of the resource, but also technicalities, media types, publication cycles, etc.

This observation seems to be especially relevant to the DDC, as its metadata will be undergoing significant changes in the near future, the switch to MARC as representation format only being the most obvious. A more subtle change is the way the concept of "editions" is reassessed to signify time-stamped snapshots of the Dewey database without wholesale changes to the referenced resources, rather than adhering to the 7-year cycle of the print edition. This conceptual change is significant to facilitate contiguous ranges of historic versions for individual concepts that can be identified and exposed for retrieval systems (Tennis, 2006).

A second (more moderate) position mandates to include only "well-behaved" metadata that is functionally dependent on the Web document, for example, is unlikely to change independently of the identified resource. In case such metadata changes, it would automatically describe a new document that in turn justifies a new URI.

On the other end of the spectrum are axioms put forward by Roy Fielding's REST (Representational State Transfer) paradigm. He states in his seminal work that it must at least be possible to *treat* URIs as opaque or mere identifiers when dereferenced. Yet the URI is most importantly a resource identifier, not a document identifier.

> [A]uthors need an identifier that closely matches the semantics they intend by a hypermedia reference, allowing the reference to remain static even though the result of accessing that reference may change over time. [… REST is] defining a resource to be the semantics of what the author intends to identify, rather than the value corresponding to those semantics at the time the reference is created. (Fielding, 2000)

This slight redefinition fits into the REST framework that aims at using URIs to actively expose and manipulate resources and their states.

While Berners-Lee emphasizes the character of the URI as a rigid and arbitrary designator, the second position concentrates on it being a locator of documents on a network, and only the third position frames the URI as a concept that allows its representations to be accessed and manipulated in various ways. In addition, RESTful URIs are considered representation-agnostic, so the way in which the data is presented will not interfere with the semantics that govern the identification of a resource.

3. URIs for the Dewey Decimal Classification

When Andy Houghton and colleagues from OCLC's Office of Research started designing a URI structure for the DDC, the result was a very elegant URI Template:

`http://dewey.info/{aspect}/{object}/{locale}/{type}/{version}/{resource}`[64]

Examples of identifiers generated by this template include `http://dewey.info/concept/338.4/en/edn/22` that retrieves or identifies the 338.4 concept in the English version of edition 22. These URIs have some very distinct advantages in being clearly structured, hackable, and (almost) entirely derivable from existing metadata, among others. They also had some drawbacks, however, in being very closely tied to a specific entity-relationship representation of DDC's conceptual structure, and based on an early draft of the URI Template specification that didn't allow for much flexibility in specifying optional and mandatory elements; e.g., segments in the path could not be skipped, only successively omitted starting from the last element. Removing an element in that manner widens the information context of the identifier (determined by the data model that was used to establish the sequence).

From a services perspective, however, this approach seems not flexible enough in the way it mandates what pieces of information agents have to possess in order to interact with the exposed resource. The identifier does not need to be an exact mapping of the data structure of the whole classification; it rather should encourage multiple views on a resource.

The feedback we have received based on the original proposal suggests that the Dewey number, even if semantically not unproblematic, should be the central part of the URI structure. Furthermore, assuming that Dewey concepts, identified by their class number, ought to have the same intension across translations, locale or language could be removed from the concept identifier altogether and handled like any other representation variant. Thirdly, thinking from a

[64] The value set of the {aspect} associated with an {object} contains at least "concept", "scheme", and "index"; {object} is a type of {aspect}, {locale} identifies a Dewey translation, {type} identifies a Dewey edition type and contains, at a minimum, the values "edn" or "abr", {version} identifies a Dewey edition version, {resource} identifies a resource associated with an {object} in the context of {locale}, {type}, and {version}.

REST perspective, identifying resources is closely interrelated to the conception of a service architecture that answers an agent's questions about those resources.

It should not be a prerequisite to already have a clear conception about the versioning conventions of DDC concepts. If we redefine editions as being nothing more than named time slices, opaque version labels assigned to a group of resources at a specific point in time, the hierarchical `{edition_type}/{edition_version}`, e.g., `edn/22`, `abr/14`, should be represented together in a more generic way as {edition_stamp} with a larger value set ("e22" for the full edition 22, "a14" for the abridged edition 14, "qr-3-2007" for the third quarterly release in 2007 of the Dewey database, or "[2007, 05, 25]" for a specific point in time that would be mapped to the most appropriate version by the service).[65]

Evolving the "edition_type" aspect to a timestamp aspect is useful on another level. With "edition_stamp" becoming just a different moniker for "time", it can be handled as yet another representational variant of a resource, alongside the representation format specified by HTTP Content-Type.

Following a similar strategy, if Dewey classes have stable intensions independently of language instantiation, the language should be handled in a similar fashion as well. Just like format as the third dimension in which a representation can vary (SKOS, MARC, HTML, etc.), the language/locale element becomes either part of the configuration of the service, query string parameter, or content negotiation. (After abstracting out language, format, and time, we arrive at what is often called a "generic resource" [Berners-Lee, 2000], addressed below in more detail.)

Using the latest draft of the URI Template specification (Gregorio, Hadley, Nottingham, & Orchard, 2008), the new structure looks like this:

```
http://dewey.info/{aspect}{-opt|/|aspect}{object}{-opt|/|object}
  {-list|/|edition_stamp}{-opt|/|edition_stamp}/{-list|/|resource}
```

Let's analyze some concrete URIs generated by expanding this template:

```
http://dewey.info/class/338.4/2007/05/25/about.en.html
http://dewey.info/class/338.4/e22/about.en.html
```

The above URIs both identify or retrieve an English HTML representation of the 338.4 concept found in edition 22.

```
http://dewey.info/class/2--74-79/2007/05/25/about
http://dewey.info/class/2--74-79/about
```

Format and language of the retrieved resources will be determined by the agent, either by content-negotiation, parsing the generic resource for RDF statements indicating available variants, or using a URI of a fixed resource.

Identifiers for other entities are built accordingly by modifying {aspect} and/or {object}[66]:

[65] Depending on the implementation, it could still be necessary to keep a mechanism to distinguish full and abridged versions independently of how their respective editorial state is labeled, for example `{edition_type}/{edition_stamp}` with {edition_type} being either "abridged" or "full", and {edition_stamp} similar as explicated above.

[66] Besides "class", which should only address assignable concepts, {aspect} might include at this point "manual", "index", "table", "scheme", and "id".

http://dewey.info/table/1/a14/about.en.skos
http://dewey.info/scheme/about

The first URI identifies a fixed representation of table one, the second URI is the generic identifier for the whole scheme, similar to `dcterms:DDC` defined by the DCMI metadata terms.

So far the `{resource}` has always just been `/about`, indicating a description of the concept found in the DDC. Following the REST paradigm, however, we can weave into the URIs collections of resources that are far more useful for services than just retrieving atomic concepts.[67]

http://dewey.info/class/338.4/e22/ancestors/about
http://dewey.info/class/338.4/ancestors/about.en.skos

Both URIs could be used to identity or retrieve the entire graph of the upward hierarchy of the given concept. The first, identifying a generic representation of the resource, could use content-negotiation and redirecting to HTML by default. Depending on service architecture decisions, a HTTP response code 300 (Multiple Choices) might be returned instead with RDF statements enumerating the choices. The second URI, while retrieving the superordinate concepts of all historic versions of the resource in English, includes links to the content in all other available languages (Raman, 2006).

Depending on what is identified as useful resources for a "Classify API", more application scenarios or use cases, like browsing, retrieval, or query expansion could be supported, by using `/children` (retrieving all immediate subclasses), `/siblings` (returning all coordinate classes with the same superclass, effectively providing a shortcut for a BT/NT traversal or subsequent requests for `/parent` and `/children`), `/related/about?degree=x` (providing the graph of referenced terms up to a specific degree. A `/search` resource resulting from e.g. a keyword search of a collection of all concepts in DDC 22 could be expressed in the same manner: `http://dewey.info/scheme/e22/search/about.de?kw=...`

4. Generic Resources

As the described scheme may produce several URIs that describe the same Dewey concept in somewhat different ways, it is desirable to be able to distinguish a canonical URI or representation (in this context sometimes called a "generic resource").

As discussed above, the definition of information resources is crucial to the architecture of the Web. But since anything might be identified by a URI, there has to be a way to indicate that a URI might denote something other than an information resource. As the resolution of the "httpRange-14 problem" the W3C TAG has decided that when dereferencing a URI and its resource can't be represented by a "message", i.e., identifies not an information resource or a "Web document", a HTTP response 303 (See Other) should be issued pointing to a description of the original resource. For specific ways of addressing these issues in practice, see Sauermann & Cyganiak (2008).

The question for our specific case is now: Is a Dewey class (or concept) an information resource? The Dublin Core Metadata Initiative defines it as a "set of conceptual resources", the DDC Glossary as a "group of objects," SKOS defines "conceptual resources" (a shorthand for concepts) as "units of thought."

[67] See for example (Binding & Tudhope, 2004). The authors, after evaluating different APIs for distributed KOS access, criticize the fact that most APIs mimic the data structure of the KOS too closely and don't support advanced operations like "chunking", i.e., the retrieval of a defined set of concepts with one request to the server.

Steering clear of the intricate philosophical problem (dating back to Maxwell's Demon) if a *group* of things constitutes an information resource while the things alone do not, the cited sources all suggest that concepts of a KOS should be treated as abstract objects (not as information resources). To represent that fact, the {resource} segment has been introduced into the URI to distinguish between the abstract DDC concept and a description of that concept. While `http://dewey.info/class/338.4` indentifies the concept, `http://dewey.info/class/338.4/about` identifies the information resource describing this concept. Since this last URI is designed to be representation-agnostic and provides links to more specific resources, it is in fact the generic representation of this resource.

The benefit of pointing the agent to a generic information resource before negotiating the contents of the representation is mainly semantic. By using this technique it is made clear that all descriptions of the identified resource are variants of the same representation and roughly convey the same information. The relationship of each of those resources to the generic resource is such that they specify one or more dimensions of its genericity.

For example, in our context `http://dewey.info/class/338.4/about` exemplifies an identifier for a generic resource, being *about* an abstract concept. The relationship between the URI and the representation of the resource it identifies may change over time, with respect to language and format requested. The use of the same URI will still be valid, however, because these new resources are considered more specific versions of the generic resource, and their respective relationships would be given as RDF statements about the dimension they specify. On the other hand, the resource that `http://dewey.info/class/338.4/2008/04/03/about` identifies or retrieves is only time-invariant but language- and format-generic, whereas for `http://dewey.info/class/338.4/2008/04/03/about.en.skos` it is completely fixed.

It should also be noted in this context that removing the language from the concept URI implies that a specific language version of a DDC concept can never be addressed as an abstract concept, but only as an information resource describing the abstract (language neutral) concept.

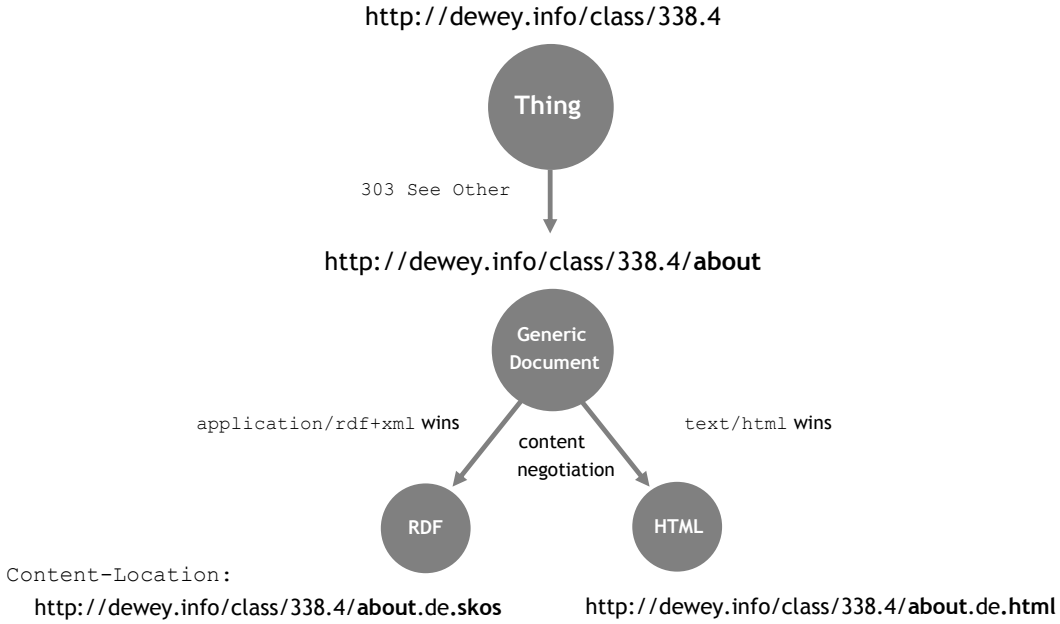

FIG. 1. Generic resources as web documents.

The concept of generic resources is especially important for designating a canonical URI for a given set of resources/representations. The findings above suggest that a candidate for a canonical URI should identify the most generic in a set of resources that can be grouped together as variants of each other.

5. Next steps

There are numerous DDC entities that have not been addressed so far and will therefore not be addressable by the URIs shown above. That doesn't mean that they won't be accessible to applications, however. Even if one assumes that these entities might be irrelevant from a service perspective, it would perhaps be useful to achieve higher granularity for users of the full Dewey data file; and in representation formats like SKOS, every reference has to be a URI, anyway. One possibility would therefore be to use opaque URIs in the `http://dewey.info/id` namespace in parallel, which, for all entities that already have other identifiers, would have to be handled as URI aliases. This set could correspond directly and exhaustively to entities in the Dewey database as represented in MARC Classification and Authorities formats, its entities could be related by OWL and even be used publicly for permalinks.

Another solution: the proposed scheme might be extended by adding fragment identifiers, enabling access to specific pieces of information beyond the level of the suggested URIs, for example, `http://dewey.info/class/1--012/e22/about#caption` to just indentify the caption "Classification" of that class, but these specific entities might be misleading if applied across different data formats (W3C Technical Architecture Group, 2004, sec. 3.2.2), e.g., MARC Classification vs. SKOS. Another potential drawback is that fragment identifiers are stripped from the URI by the user agent, so a service endpoint will never see them.

The usefulness of "shortcuts" has to be addressed in general as well. Every time a "default" is introduced, the expressiveness of the scheme is impoverished by *de facto* defining URI aliases for some resources. If `http://dewey.info/concept/338.4` defaults today (using my current Web browser) to the same representation that is retrieved by `http://dewey.info/concept/338.4/2008/04/04/about.en.html`, the possibility is lost to use the original URI as a canonical identifier for the 338.4 concept independently of time, language, or format. Yet such an identifier is a powerful tool that could retrieve all information about translations, former versions of this concept etc. as OWL or RDF expressions, making it possible for an agent to just work from this resource for any given concept. A better general way of indicating shortcuts would be to interpret an unspecified {aspect} segment as trigger for defaulting behavior, for example: only `http://dewey.info/338.4` would be defined as an alias of the fixed resource shown above, but not `http://dewey.info/concept/338.4`.

References

Berners-Lee, Tim. (2000). Web architecture: Generic resources. Retrieved April 3, 2008, from http://www.w3.org/DesignIssues/Generic.html.

Berners-Lee, Tim, Roy Fielding , and Larry Masinter. (2005). *Uniform Resource Identifier (URI): Generic Syntax*. Standard, IETF. Retrieved from http://www.ietf.org/rfc/rfc3986.txt.

Berrueta, Diego, and Jon Phipps (2008). Best practice recipes for publishing RDF vocabularies. *Working Draft, W3C.* Retrieved from http://www.w3.org/TR/2008/WD-swbp-vocab-pub-20080123/.

Binding, Ceri, and Douglas Tudhope. (2004). KOS at your service: Programmatic access to knowledge organisation systems. *Journal of Digital Information, 4*(4). Retrieved from http://journals.tdl.org/jodi/article/view/jodi-124/109.

Coyle, Karen, and Diane Hillmann. (2007). Resource Description and Access (RDA): Cataloging rules for the 20th century. *D-Lib Magazine, 13*(1/2). Retrieved from http://dlib.org/dlib/january07/coyle/01coyle.html.

Fielding, Roy (2000). *Architectural styles and the design of network-based software architectures.* Thesis (Ph. D., Information and Computer Science), University of California, Irvine. Retrieved from http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm.

Fielding, Roy, and Ian Jacobs. (2006). Authoritative metadata. *TAG Finding, W3C*. Retrieved from http://www.w3.org/2001/tag/doc/mime-respect.html.

Gregorio, Joe, M. Hadley, M. Nottingham, and D. Orchard. (2008). URI Template. *Internet-Draft, IETF*. Retrieved from http://www.ietf.org/internet-drafts/draft-gregorio-uritemplate-03.txt.

Mendelsohn, Noah, and Stuart Williams. (2007). The use of metadata in URIs. *TAG Finding, W3C*. Retrieved from http://www.w3.org/2001/tag/doc/metaDataInURI-31-20070102.html.

Raman, T. V. (2006). On Linking Alternative Representations To Enable Discovery And Publishing. *TAG Finding, W3C.* Retrieved from http://www.w3.org/2001/tag/doc/alternatives-discovery-20061101.html.

Sauermann, Leo, and Richard Cyganiak. (2008). Cool URIs for the semantic web. *Working Draft, W3C*. Retrieved from http://www.w3.org/TR/2008/WD-cooluris-20080321/.

Tennis, Joseph T. (2006). Versioning concept schemes for persistent retrieval. *Bulletin of the American Society for Information Science and Technology, 32*(5), 13–16. doi: 10.1002/bult.2006.1720320506

W3C Technical Architecture Group. (2004). Architecture of the world wide web, volume one. *Recommendation, W3C*. Retrieved from http://www.w3.org/TR/2004/REC-webarch-20041215/.